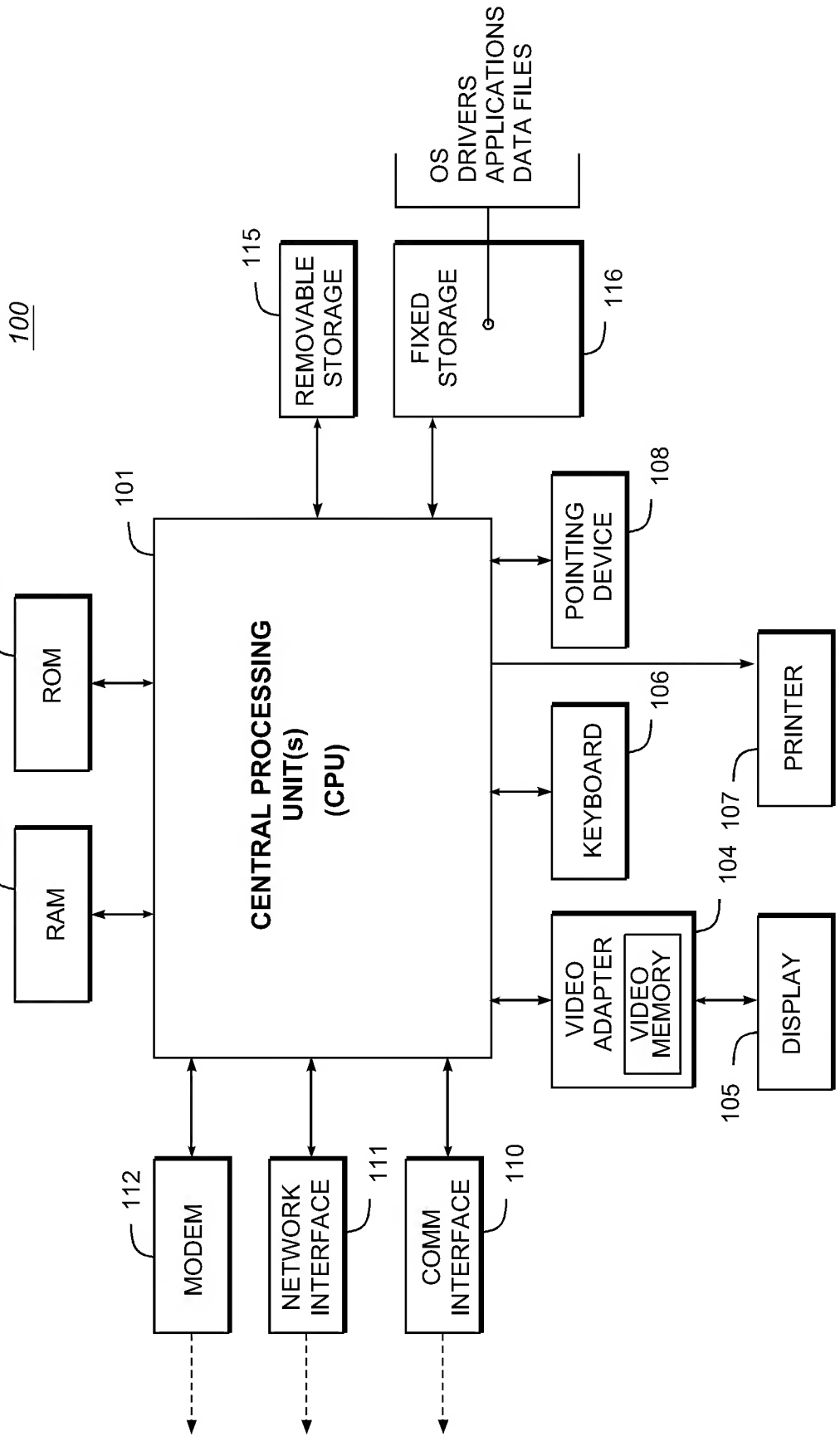

FIG. 1
(PRIOR ART)

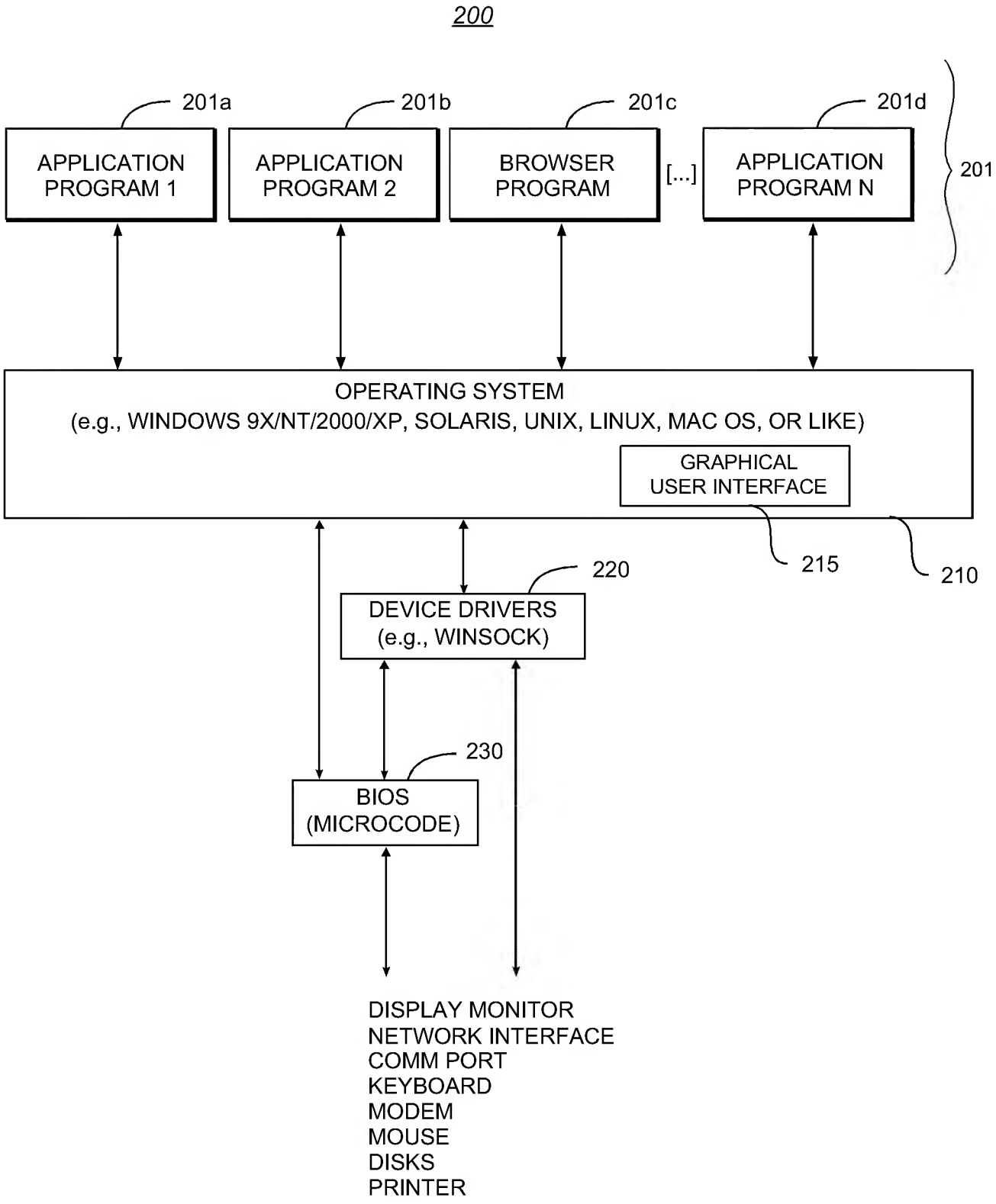
200
201a
201b
201c
201d
APPLICATION PROGRAM 1
APPLICATION PROGRAM 2
BROWSER PROGRAM
[...]
APPLICATION PROGRAM N
201
OPERATING SYSTEM
(e.g., WINDOWS 9X/NT/2000/XP, SOLARIS, UNIX, LINUX, MAC OS, OR LIKE)
GRAPHICAL USER INTERFACE
215
210
220
DEVICE DRIVERS
(e.g., WINSOCK)
230
BIOS
(MICROCODE)
DISPLAY MONITOR
NETWORK INTERFACE
COMM PORT
KEYBOARD
MODEM
MOUSE
DISKS
PRINTER

FIG. 2

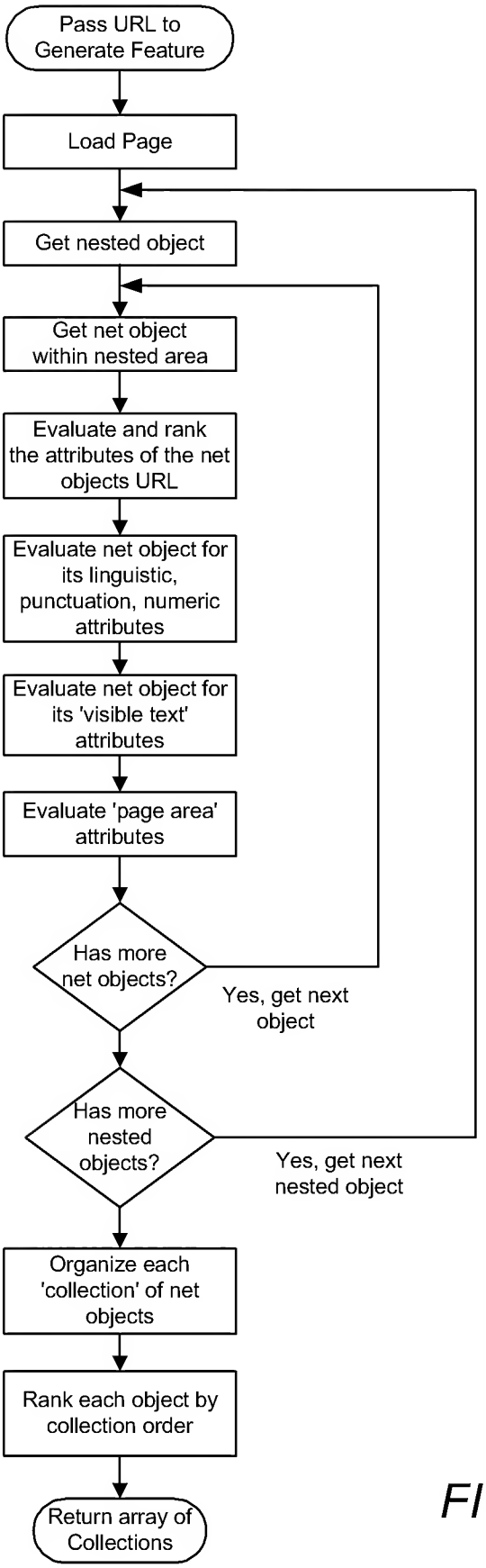
Pass URL to Generate Feature
300
Load Page
302
Get nested object
304
Get net object within nested area
Evaluate and rank the attributes of the net objects URL
Evaluate net object for its linguistic, punctuation, numeric attributes
Evaluate net object for its 'visible text' attributes
Evaluate 'page area' attributes
Has more net objects?
Yes, get next object
306
Has more nested objects?
Yes, get next nested object
308
Organize each 'collection' of net objects
Rank each object by collection order
Return array of Collections

FIG. 3

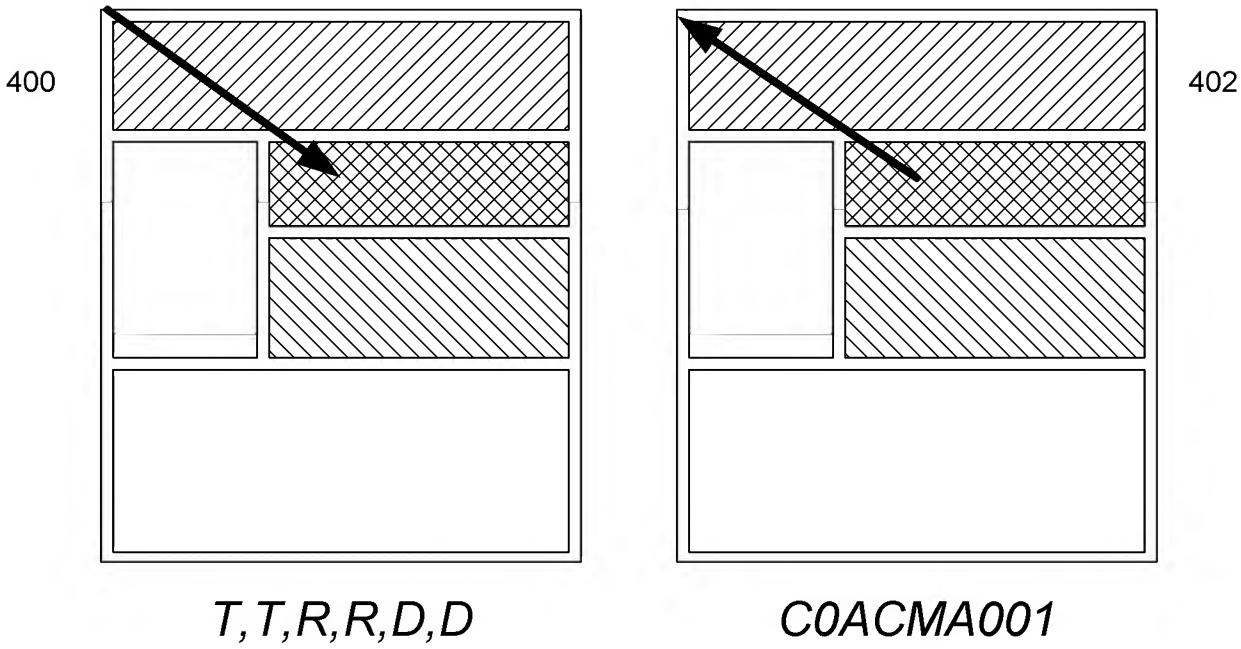

FIG. 4

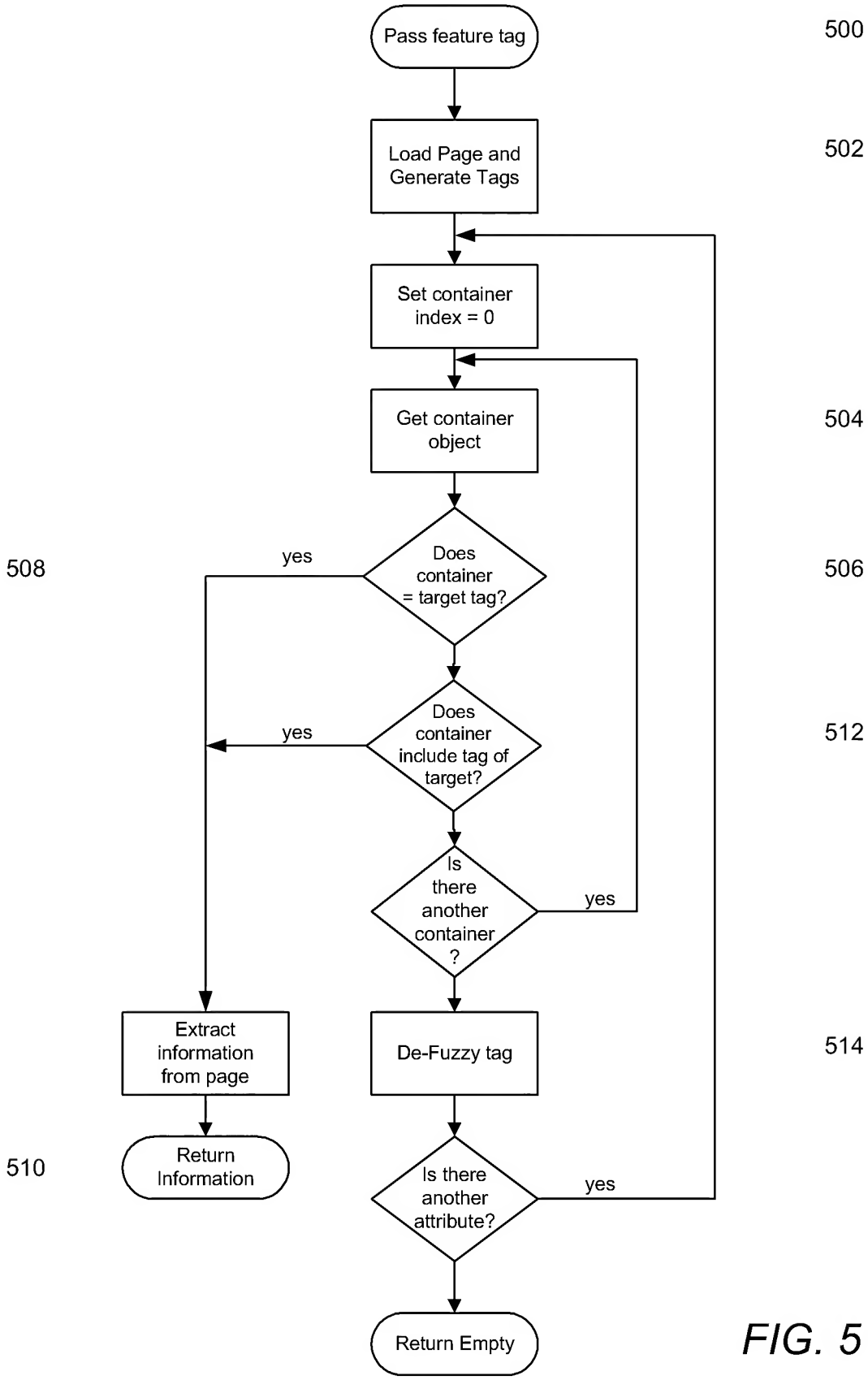
Pass feature tag
500
Load Page and Generate Tags
502
Set container index = 0
Get container object
504
Does container = target tag?
506
yes
508
Does container include tag of target?
yes
512
Is there another container ?
yes
Extract information from page
De-Fuzzy tag
514
Return Information
510
Is there another attribute?
yes
Return Empty

FIG. 5

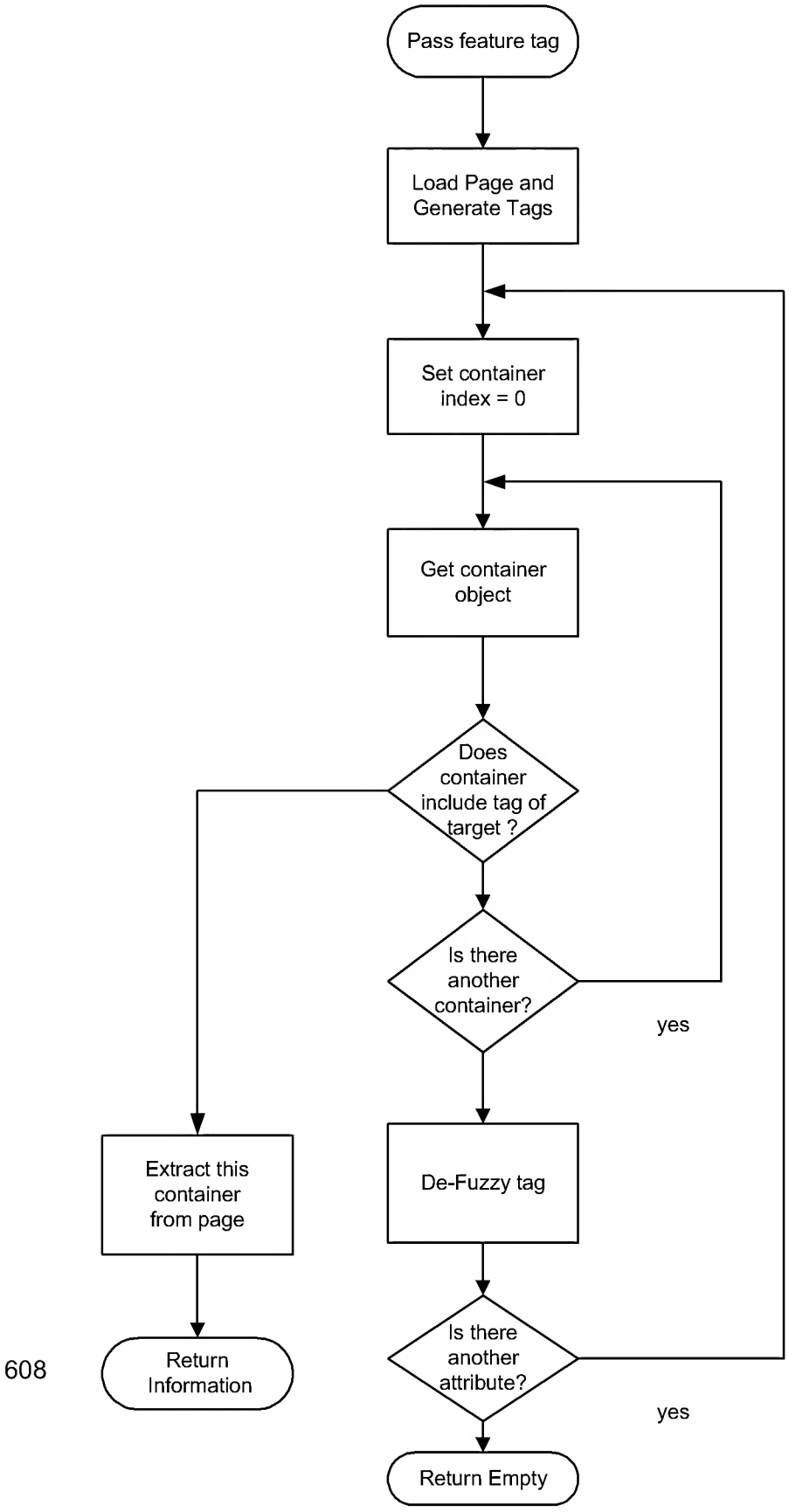
Pass feature tag
600
Load Page and Generate Tags
602
Set container index = 0
Get container object
604
Does container include tag of target ?
606
Is there another container?
yes
Extract this container from page
De-Fuzzy tag
Is there another attribute?
yes
608
Return Information
Return Empty

FIG. 6

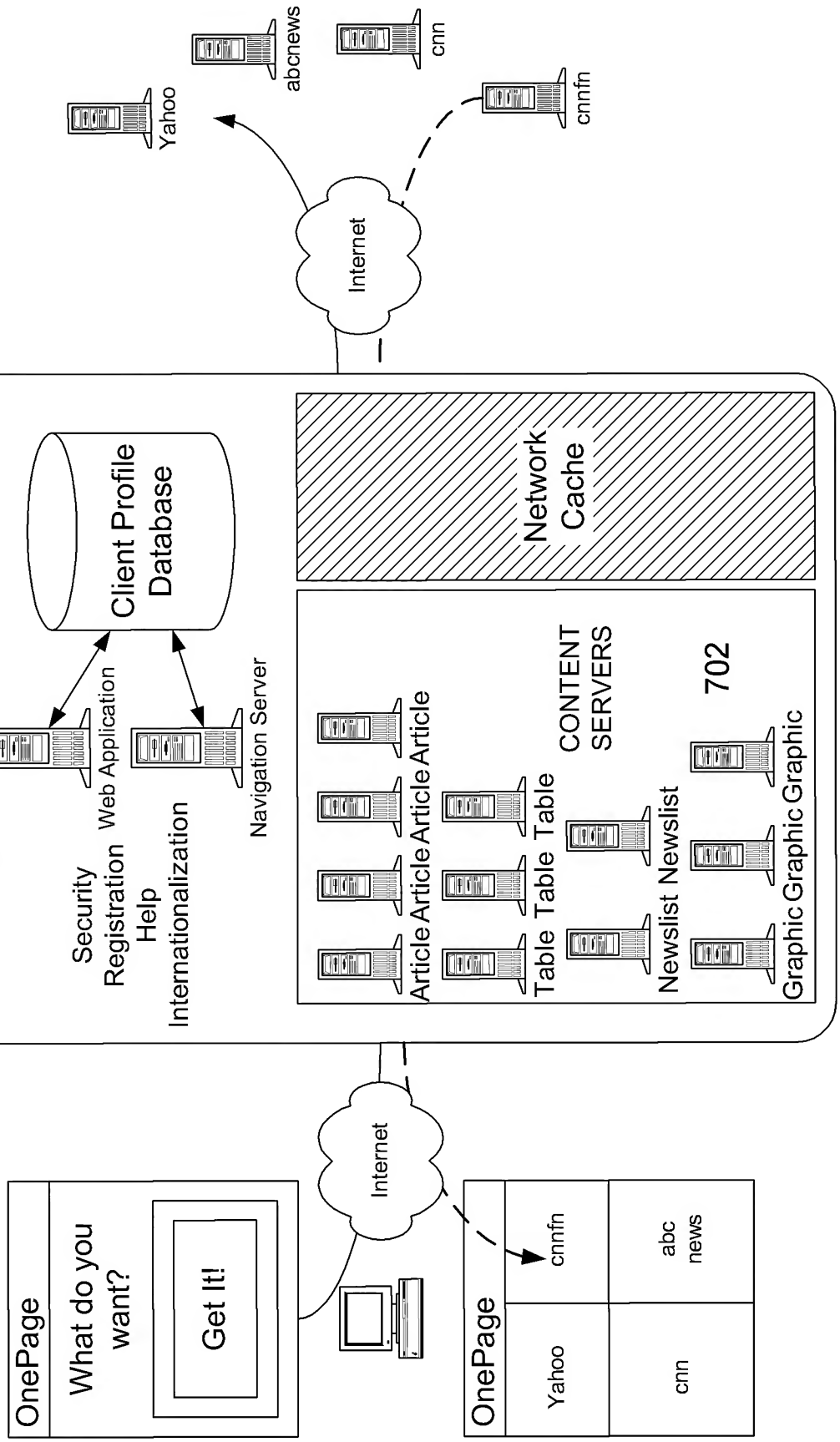
OnePage
What do you want?
Get It!
Internet
OnePage
Yahoo
cnnfn
cnn
abc news
Internet
Yahoo
abcnews
cnn
cnnfn
700
OnePage
Security
Registration
Help
Internationalization
Web Application
Navigation Server
Client Profile Database
Network Cache
CONTENT SERVERS
Article
Article
Article
Article
Table
Table
Table
Newslist
Newslist
Graphic
Graphic
Graphic
702

FIG. 7

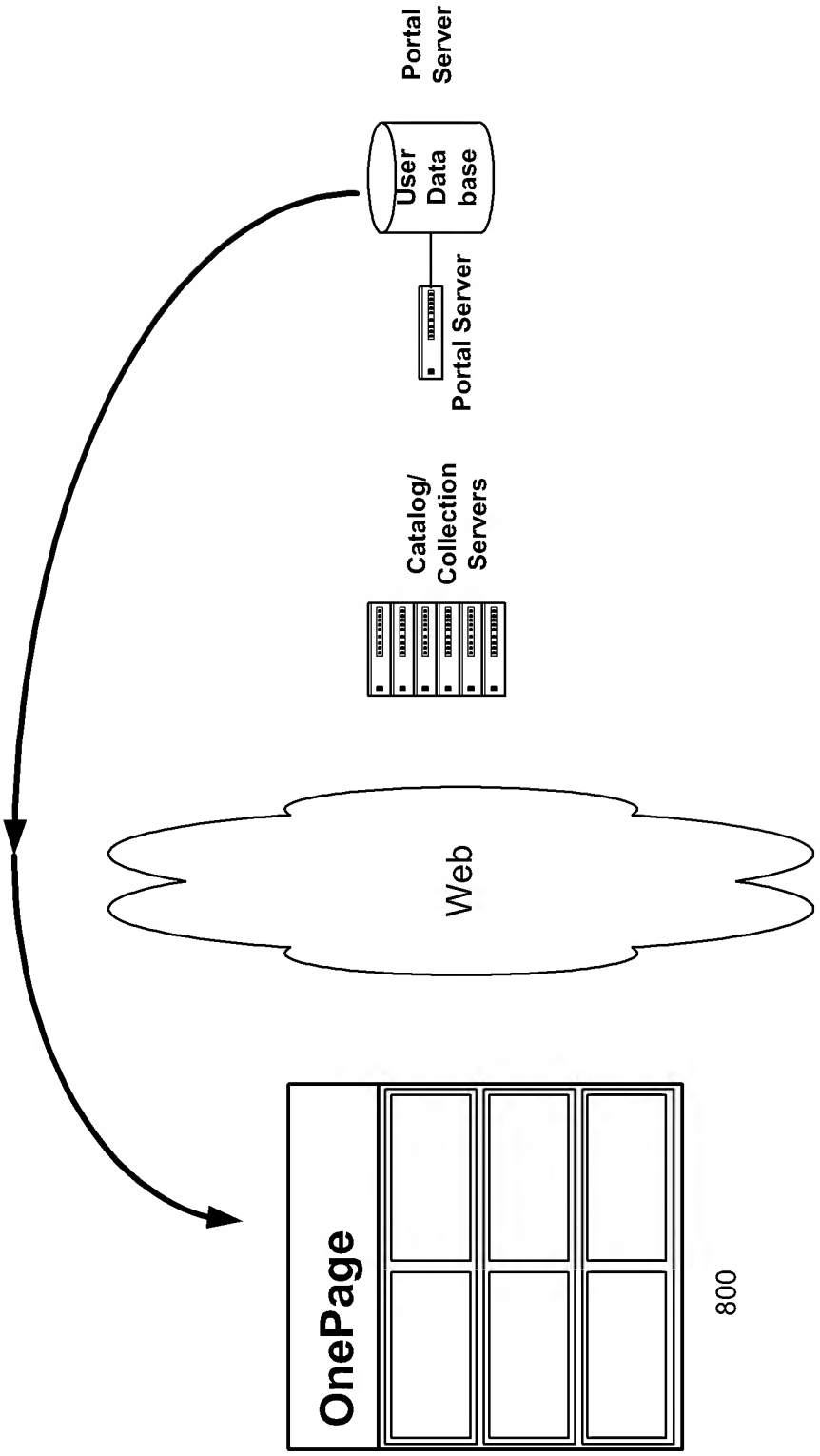
Portal Server
User Data base
Portal Server
Catalog/ Collection Servers
Web
OnePage
800

FIG. 8A

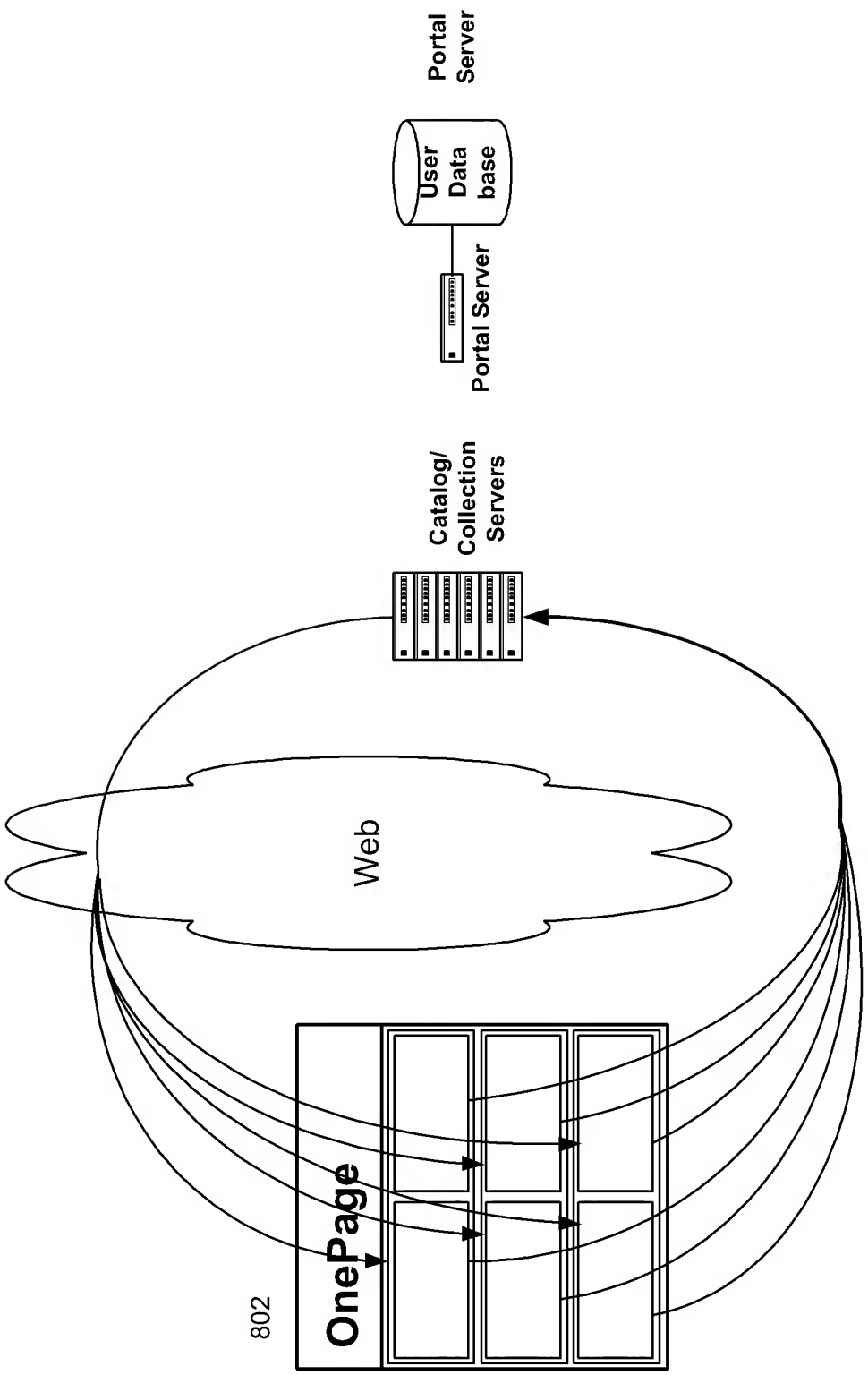
Portal Server
User Data base
Portal Server
Catalog/ Collection Servers
Web
OnePage
802

FIG. 8B

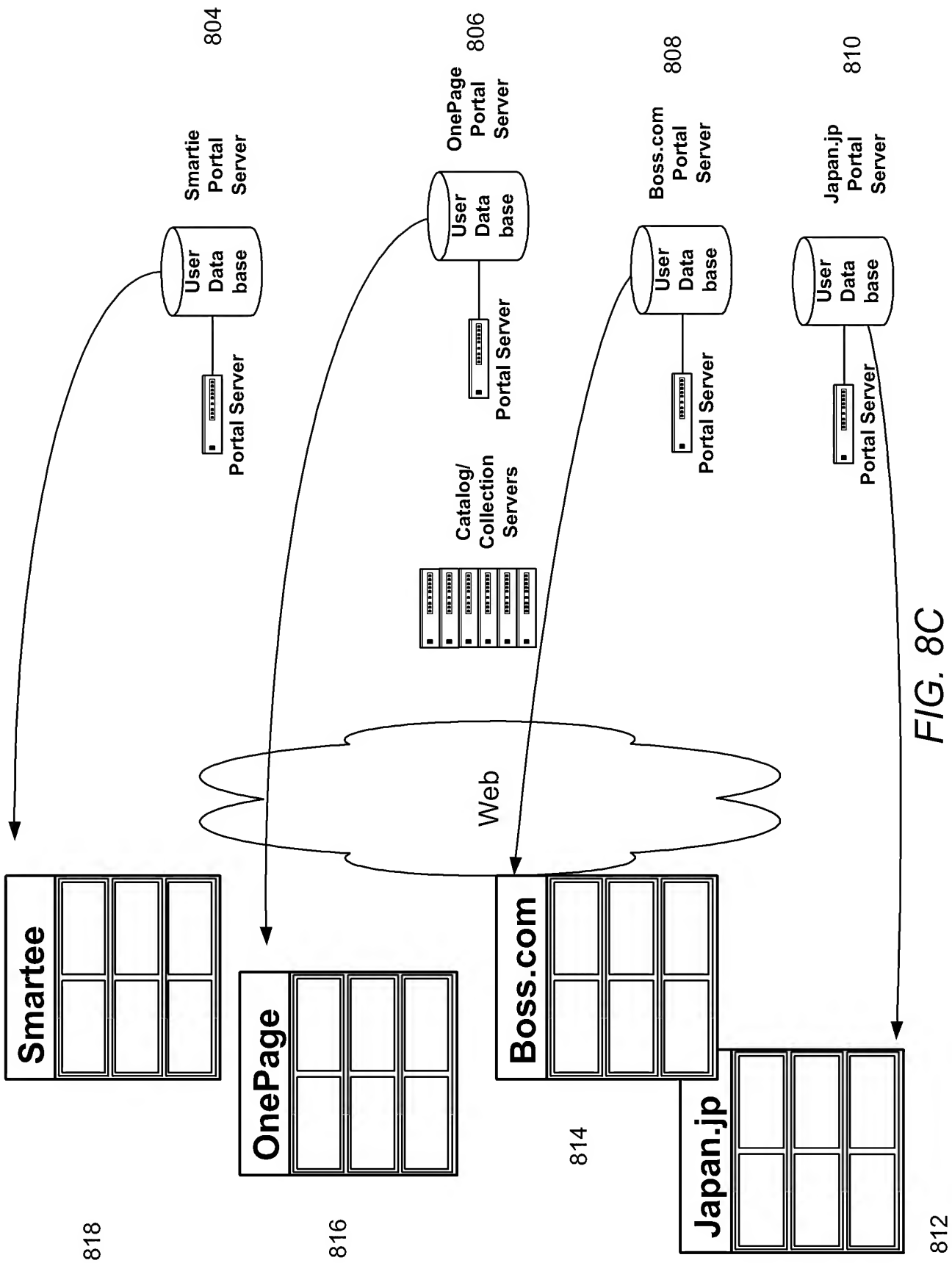
804
Smartie Portal Server
User Data base
Portal Server
806
OnePage Portal Server
User Data base
Portal Server
808
Boss.com Portal Server
User Data base
Portal Server
810
Japan.jp Portal Server
User Data base
Portal Server
Catalog/ Collection Servers
Web
Smartee
OnePage
Boss.com
Japan.jp
818
816
814
812

FIG. 8C

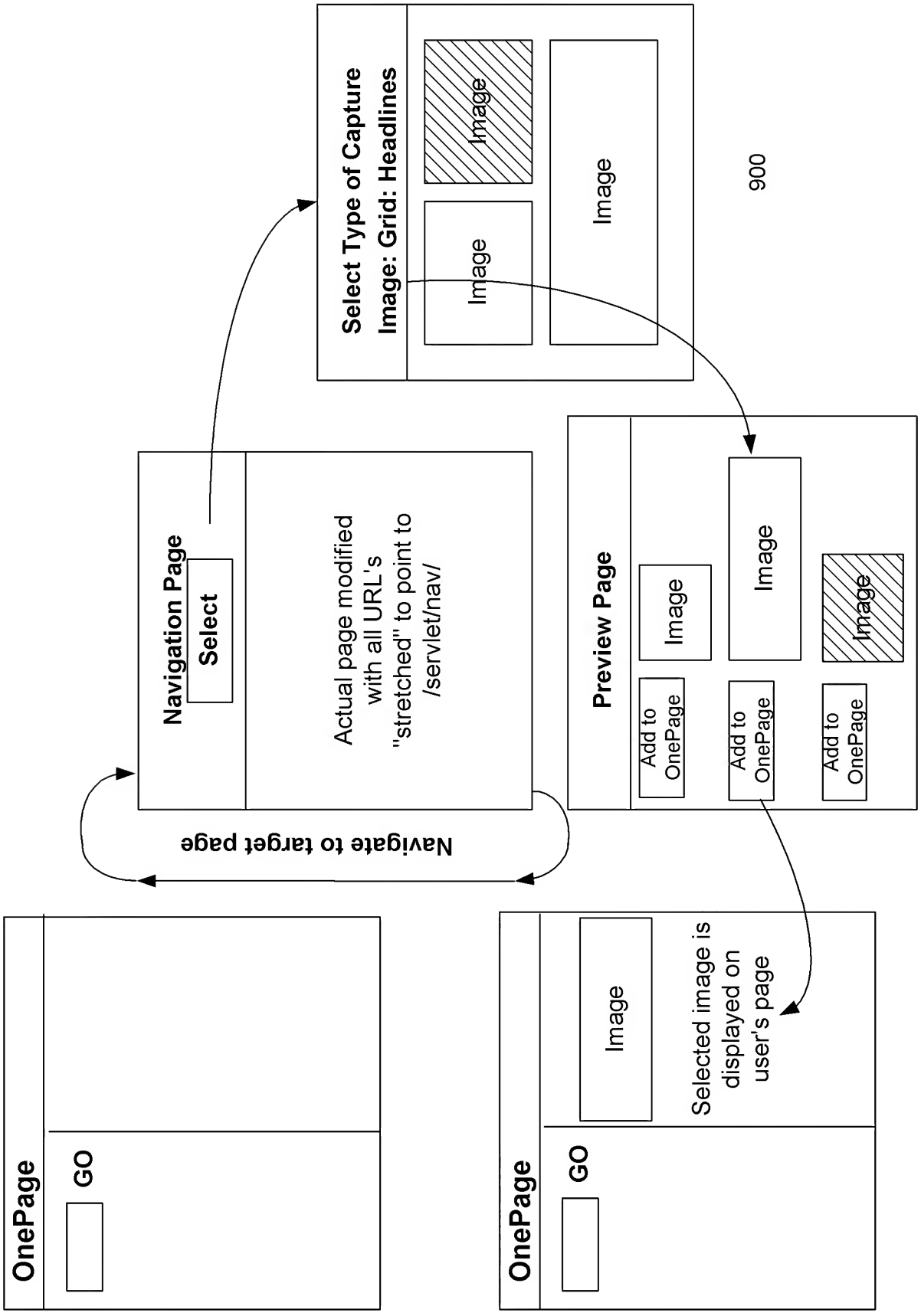
OnePage
GO
Navigate to target page
Navigation Page
Select
Actual page modified with all URL's "stretched" to point to /servlet/nav/
Select Type of Capture
Image: Grid: Headlines
Image
Image
Image
900
Preview Page
Add to OnePage
Image
Add to OnePage
Image
Add to OnePage
Image
OnePage
GO
Image
Selected image is displayed on user's page

FIG. 9

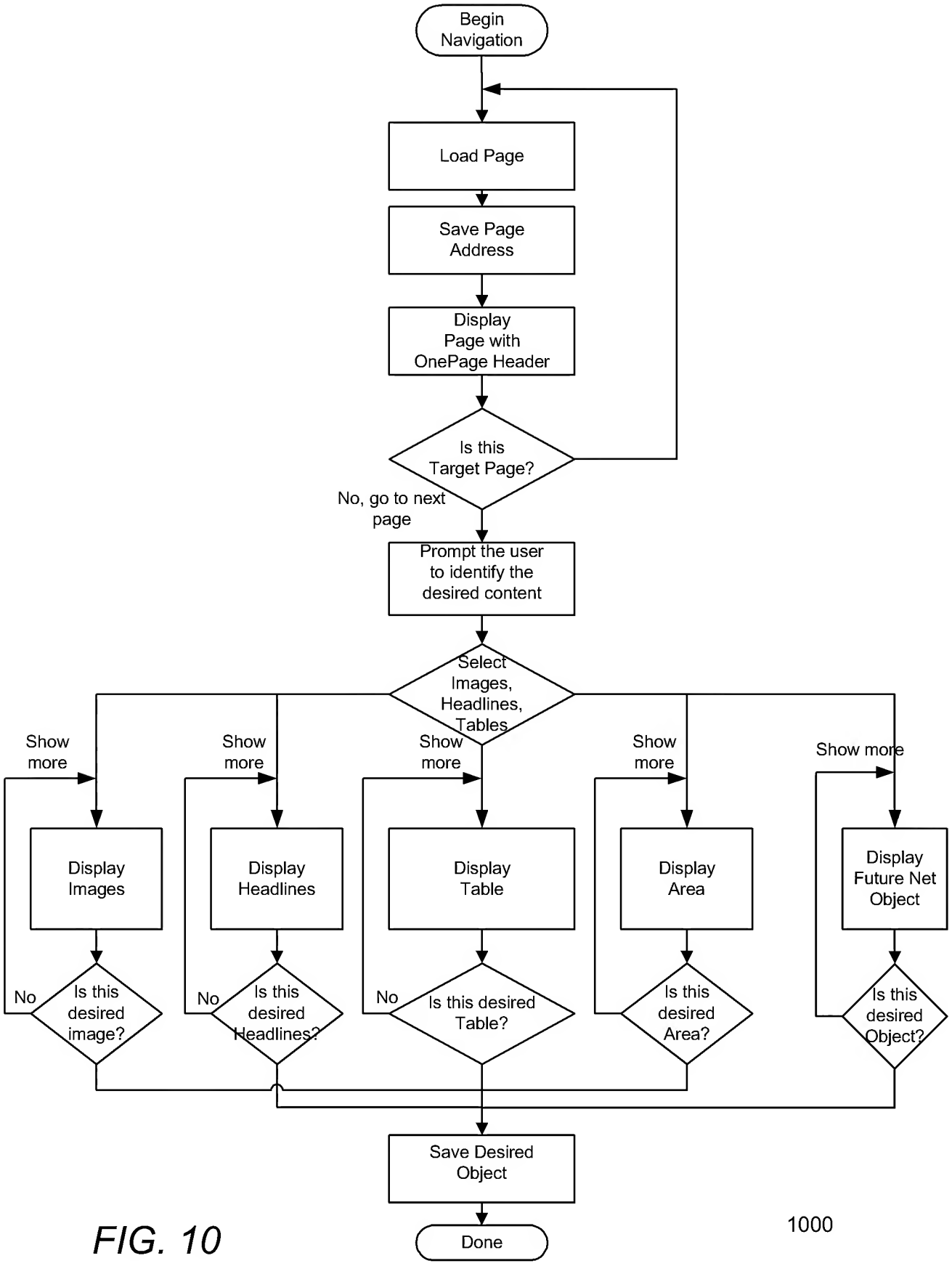
Begin Navigation
Load Page
Save Page Address
Display Page with OnePage Header
Is this Target Page?
No, go to next page
Prompt the user to identify the desired content
Select Images, Headlines, Tables
Show more
Display Images
Is this desired image?
No
Show more
Display Headlines
Is this desired Headlines?
No
Show more
Display Table
Is this desired Table?
No
Show more
Display Area
Is this desired Area?
Show more
Display Future Net Object
Is this desired Object?
Save Desired Object
Done
1000

FIG. 10

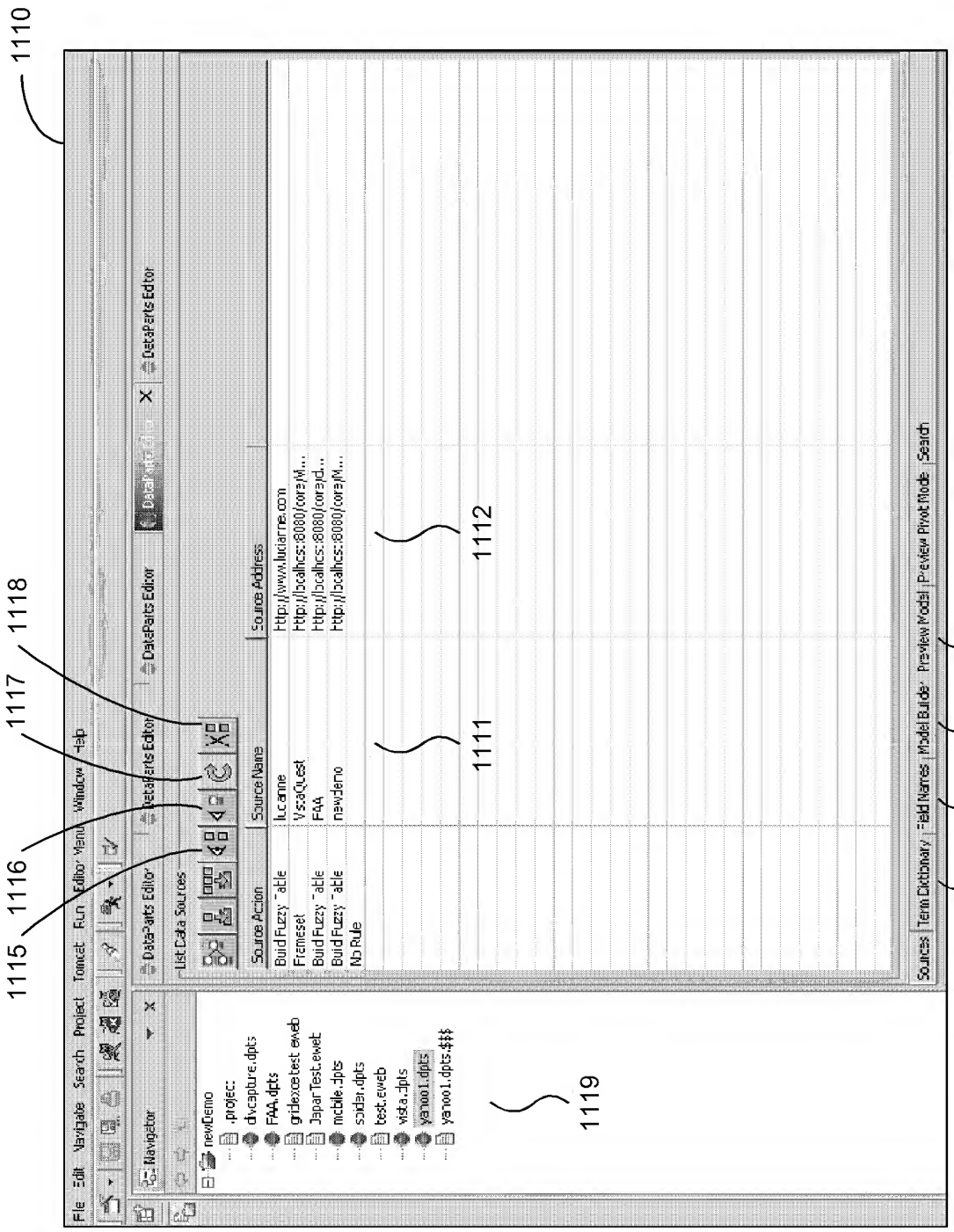
1110
1115
1116
1117
1118
1119
1111
1112
1121
1122
1123
1124
File Edit Navigate Search Project Tomcat Run Editor Menu Window Help
Navigator
newDemo
.project
dvcapture.dpts
FAA.dpts
gridexcetest.eweb
JapanTest.eweb
mobile.dpts
sonar.dpts
test.eweb
vista.dpts
yahoo1.dpts
yahoo1.dpts:$$$
DataParts Editor
List Data Sources
Source Action
Source Name
Source Address
Build Fuzzy Table
lucianne
http://www.lucianne.com
Frameset
VSeaQuest
http://localhost:8080/corsa/M...
Build Fuzzy Table
FAA
http://localhost:8080/corsa/d...
Build Fuzzy Table
newdemo
http://localhost:8080/corsa/M...
No Rule
Sources
Term Dictionary
Field Names
Model Builder
Preview Model
Preview Pivot Mode
Search

FIG. 11A

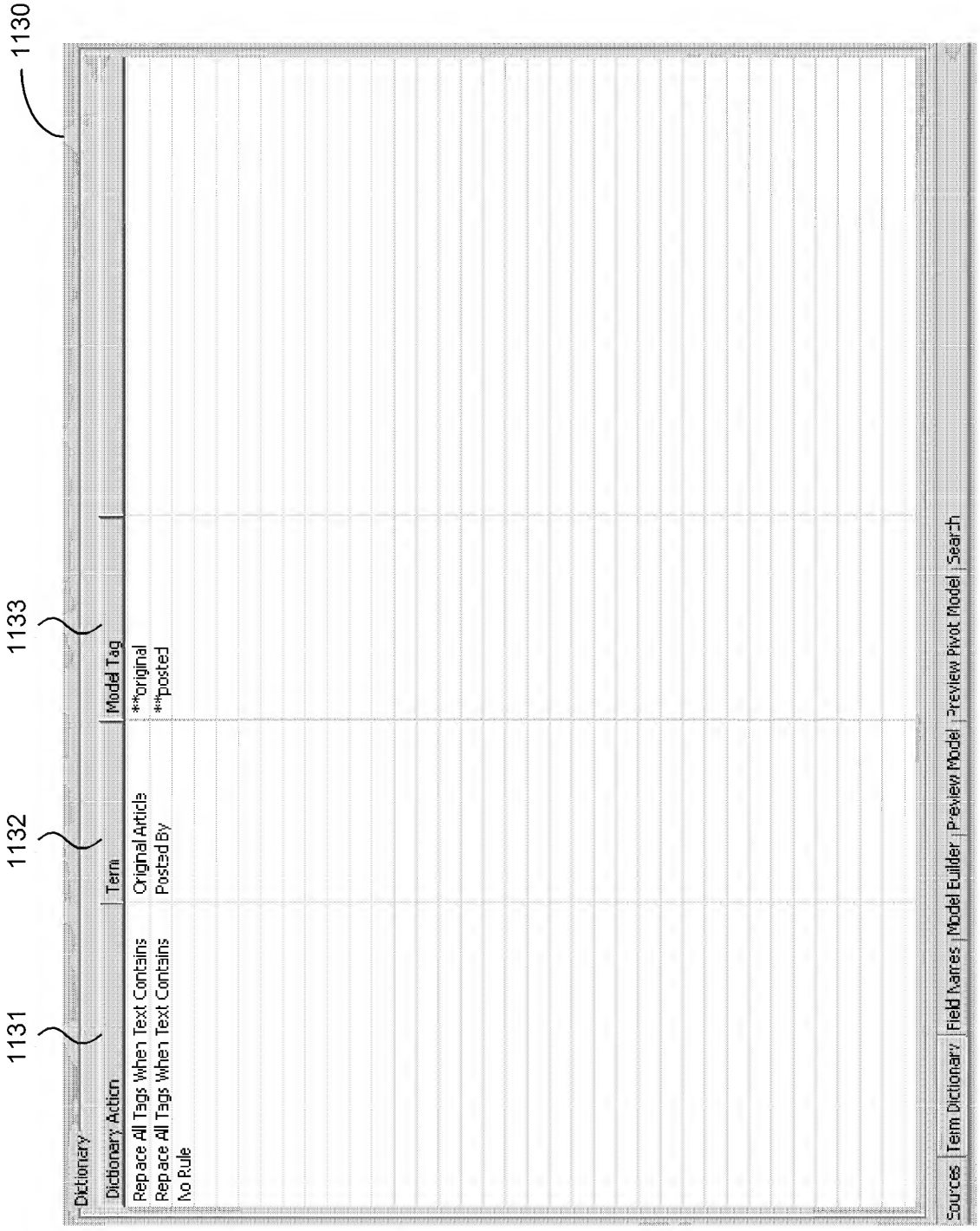
1130
1131
1132
1133
Dictionary
Dictionary Action
Term
Model Tag
Replace All Tags When Text Contains
Original Article
**original
Replace All Tags When Text Contains
Posted By
**posted
No Rule
Sources
Term Dictionary
Field Names
Model Builder
Preview Model
Preview Pivot Model
Search

FIG. 11B

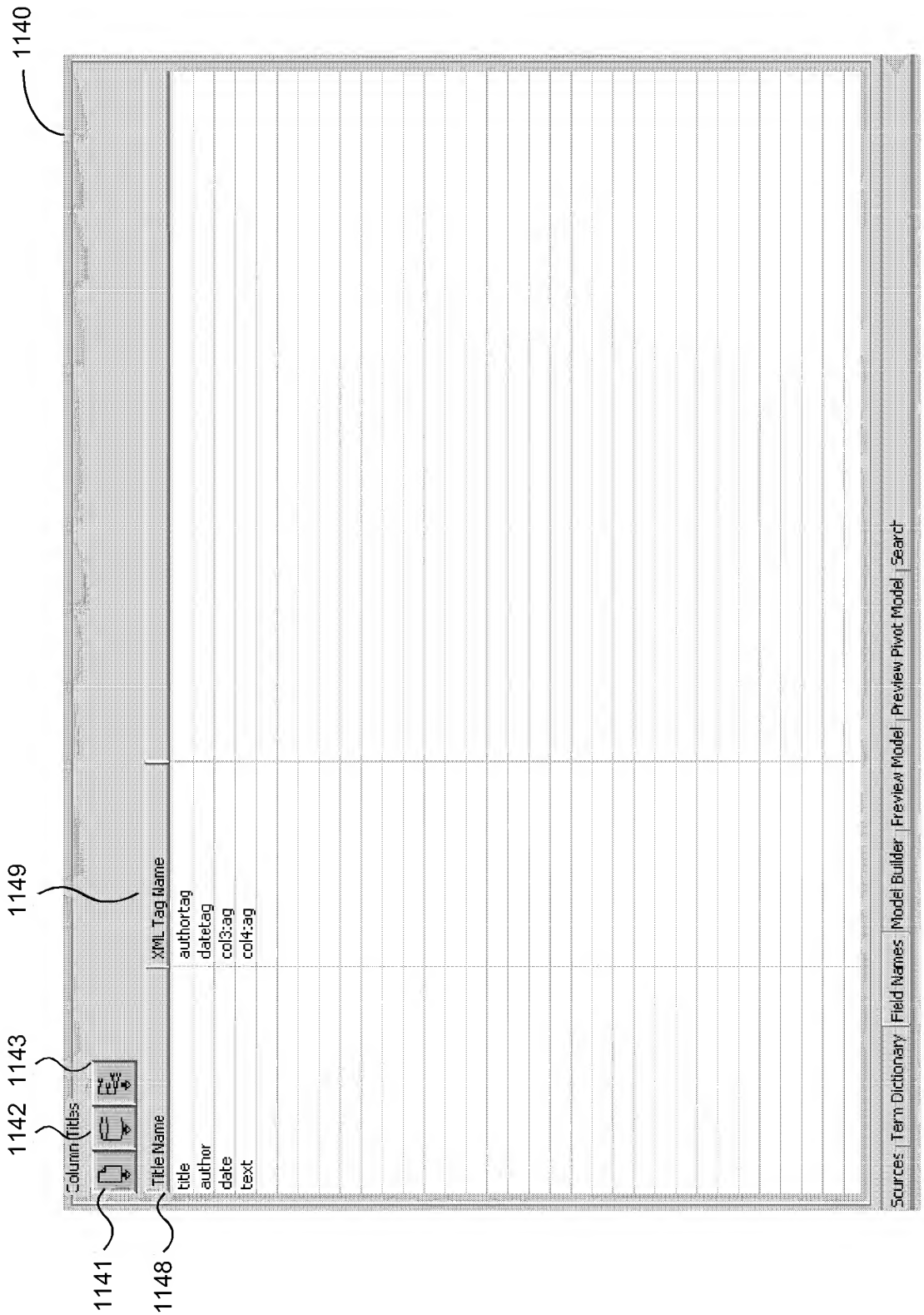
1140
1149
1142
1143
1141
1148
Column Titles
Title Name
XML Tag Name
title
author
date
text
authortag
datetag
col3:ag
col4:ag
Sources
Term Dictionary
Field Names
Model Builder
Preview Model
Preview Pivot Model
Search

FIG. 11C

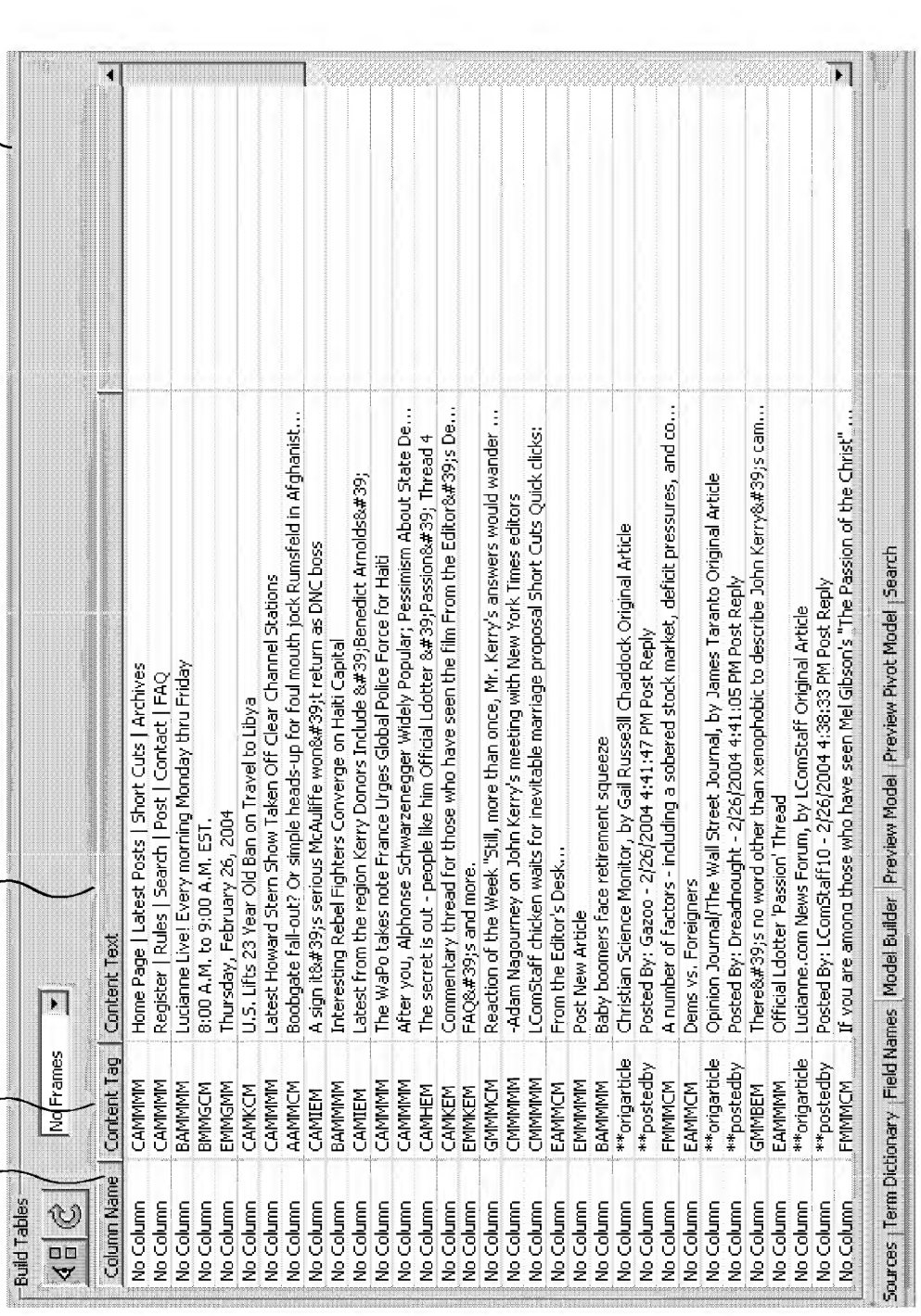
1150
1151
1152
1153
Build Tables
No Frames
Column Name
Content Tag
Content Text
No Column CAMMMM Home Page | Latest Posts | Short Cuts | Archives
No Column CAMMMM Register | Rules | Search | Post | Contact | FAQ
No Column BAMMMM Lucianne Live! Every morning Monday thru Friday
No Column BMMGCM 8:00 A.M. to 9:00 A.M. EST.
No Column EMMGMM Thursday, February 26, 2004
No Column CAMKCM U.S. Lifts 23 Year Old Ban on Travel to Libya
No Column CAMMMM Latest Howard Stern Show Taken Off Clear Channel Stations
No Column AAMMCM Boobgate fall-out? Or simple heads-up for foul mouth jock Rumsfeld in Afghanist...
No Column CAMIEM A sign it's serious McAuliffe won't return as DNC boss
No Column BAMMMM Interesting Rebel Fighters Converge on Haiti Capital
No Column CAMIEM Latest from the region Kerry Donors Include 'Benedict Arnolds'
No Column CAMMMM The WaPo takes note France Urges Global Police Force for Haiti
No Column CAMMMM After you, Alphonse Schwarzenegger Widely Popular; Pessimism About State De...
No Column CAMHEM The secret is out - people like him Official Ldotter 'Passion' Thread 4
No Column CAMKEM Commentary thread for those who have seen the film From the Editor's De...
No Column EMMKEM FAQ's and more.
No Column GMMMCM Reaction of the Week "Still, more than once, Mr. Kerry's answers would wander ...
No Column CMMMMM -Adam Nagourney on John Kerry's meeting with New York Times editors
No Column CMMMMM LComStaff chicken waits for inevitable marriage proposal Short Cuts Quick clicks:
No Column EAMMCM From the Editor's Desk...
No Column EMMMMM Post New Article
No Column BAMMMM Baby boomers face retirement squeeze
No Column **origarticle Christian Science Monitor, by Gail Russell Chaddock Original Article
No Column **postedby Posted By: Gazoo - 2/26/2004 4:41:47 PM Post Reply
No Column FMMMCM A number of factors - including a sobered stock market, deficit pressures, and co...
No Column EAMMCM Dems vs. Foreigners
No Column **origarticle Opinion Journal/The Wall Street Journal, by James Taranto Original Article
No Column **postedby Posted By: Dreadnought - 2/26/2004 4:41:05 PM Post Reply
No Column GMMBEM There's no word other than xenophobic to describe John Kerry's cam...
No Column EAMMMM Official Ldotter 'Passion' Thread
No Column **origarticle Lucianne.com News Forum, by LComStaff Original Article
No Column **postedby Posted By: LComStaff10 - 2/26/2004 4:38:33 PM Post Reply
No Column FMMMCM If you are among those who have seen Mel Gibson's "The Passion of the Christ"...
Sources | Term Dictionary | Field Names | Model Builder | Preview Model | Preview Pivot Model | Search

FIG. 11D

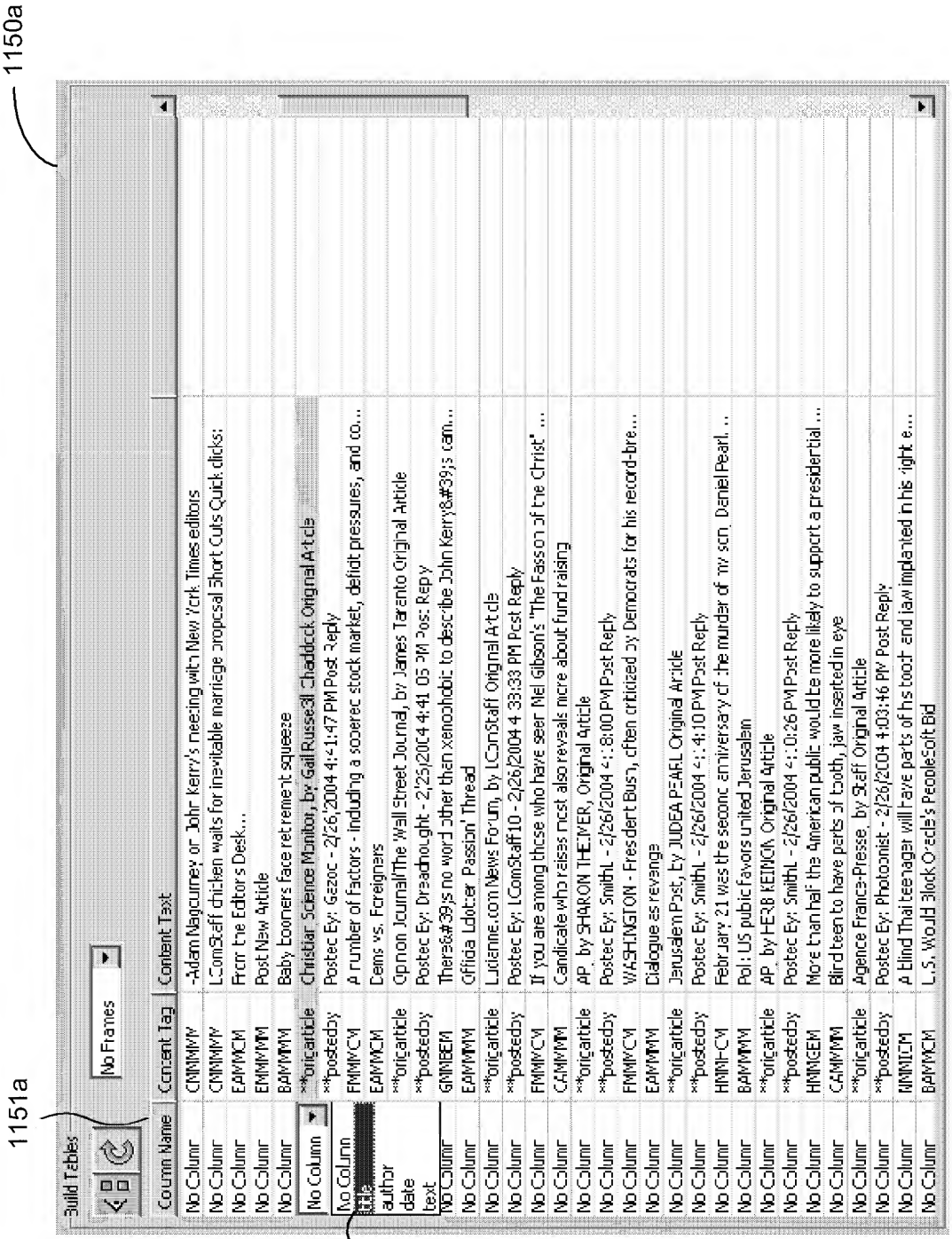
1150a
1151a
1155
Build Tables
No Frames
Column Name
Content Tag
Content Text
No Column
author
date
text
title

FIG. 11E

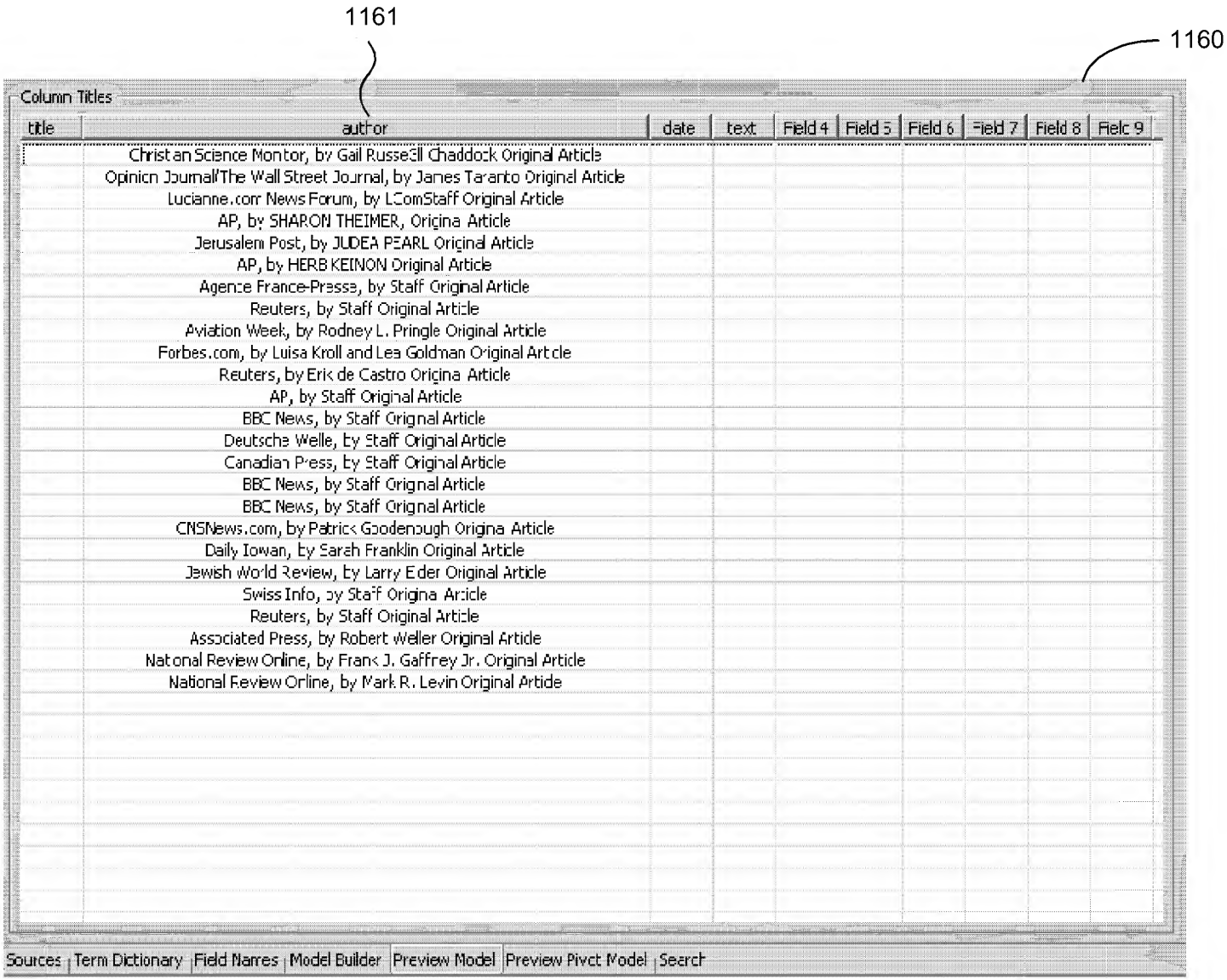
1161
1160
Column Titles
title
author
date
text
Field 4
Field 5
Field 6
Field 7
Field 8
Field 9
Christian Science Monitor, by Gail Russell Chaddock Original Article
Opinion Journal/The Wall Street Journal, by James Taranto Original Article
Lucianne.com News Forum, by LComStaff Original Article
AP, by SHARON THEIMER, Original Article
Jerusalem Post, by JUDEA PEARL Original Article
AP, by HERB KEINON Original Article
Agence France-Presse, by Staff Original Article
Reuters, by Staff Original Article
Aviation Week, by Rodney L. Pringle Original Article
Forbes.com, by Luisa Kroll and Lea Goldman Original Article
Reuters, by Eric de Castro Original Article
AP, by Staff Original Article
BBC News, by Staff Original Article
Deutsche Welle, by Staff Original Article
Canadian Press, by Staff Original Article
BBC News, by Staff Original Article
BBC News, by Staff Original Article
CNSNews.com, by Patrick Goodenough Original Article
Daily Iowan, by Sarah Franklin Original Article
Jewish World Review, by Larry Eder Original Article
Swiss Info, by Staff Original Article
Reuters, by Staff Original Article
Associated Press, by Robert Weller Original Article
National Review Online, by Frank J. Gaffney Jr. Original Article
National Review Online, by Mark R. Levin Original Article
Sources
Term Dictionary
Field Names
Model Builder
Preview Model
Preview Pivot Model
Search

FIG. 11F

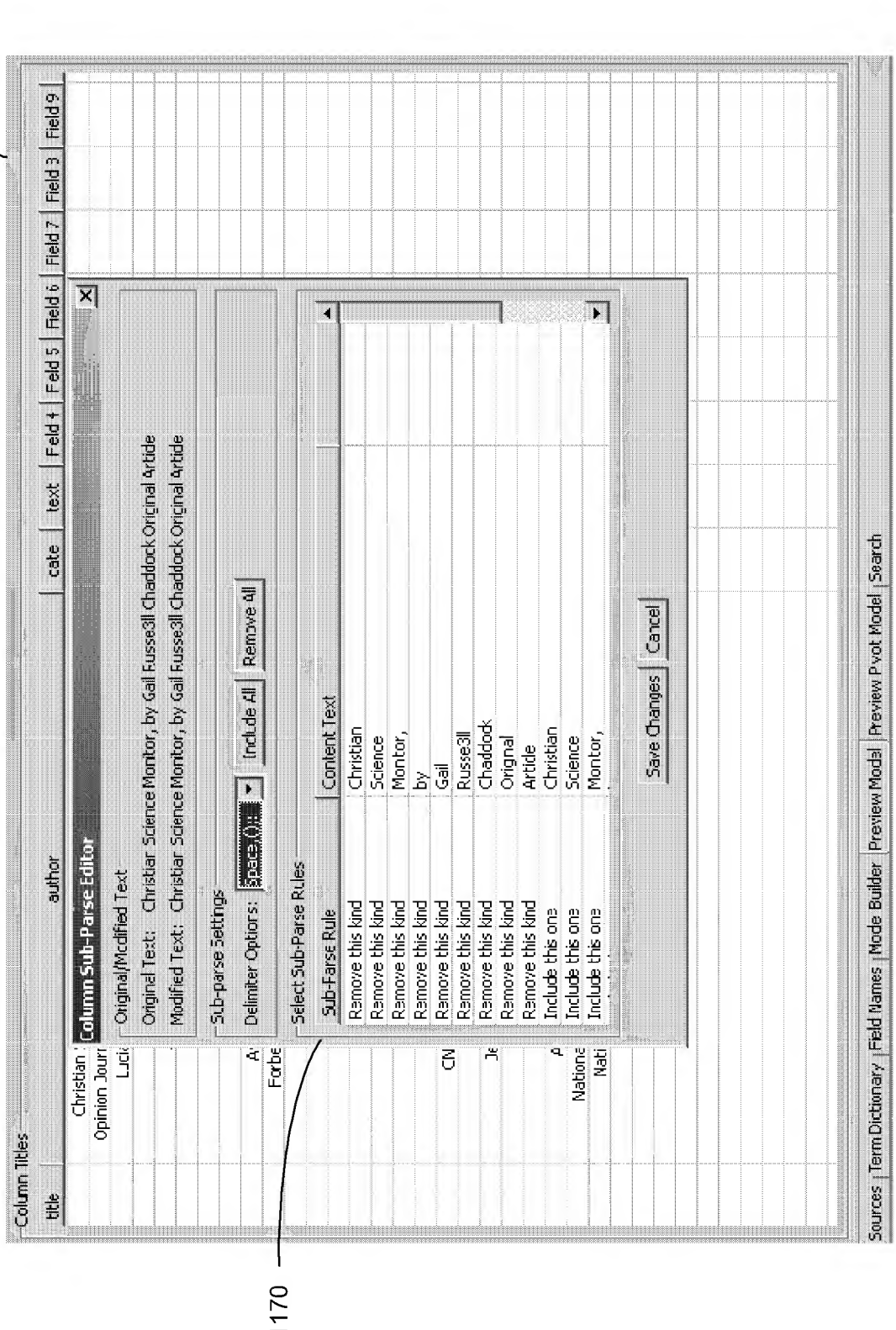
1160a
1170
Column Titles
title
author
cate
text
Field 4
Field 5
Field 6
Field 7
Field 3
Field 9
Column Sub-Parse Editor
Original/Modified Text
Original Text: Christian Science Monitor, by Gail Russe3ll Chaddock Original Artide
Modified Text: Christian Science Monitor, by Gail Russe3ll Chaddock Original Artide
Sub-parse Settings
Delimiter Options:
space ()
Include All
Remove All
Select Sub-Parse Rules
Sub-Parse Rule
Content Text
Remove this kind
Christian
Remove this kind
Science
Remove this kind
Monitor,
Remove this kind
by
Remove this kind
Gail
Remove this kind
Russe3ll
Remove this kind
Chaddock
Remove this kind
Original
Remove this kind
Artide
Include this one
Christian
Include this one
Science
Include this one
Monitor,
Save Changes
Cancel
Sources
Term Dictionary
Field Names
Mode Builder
Preview Model
Preview Pivot Model
Search

FIG. 11G

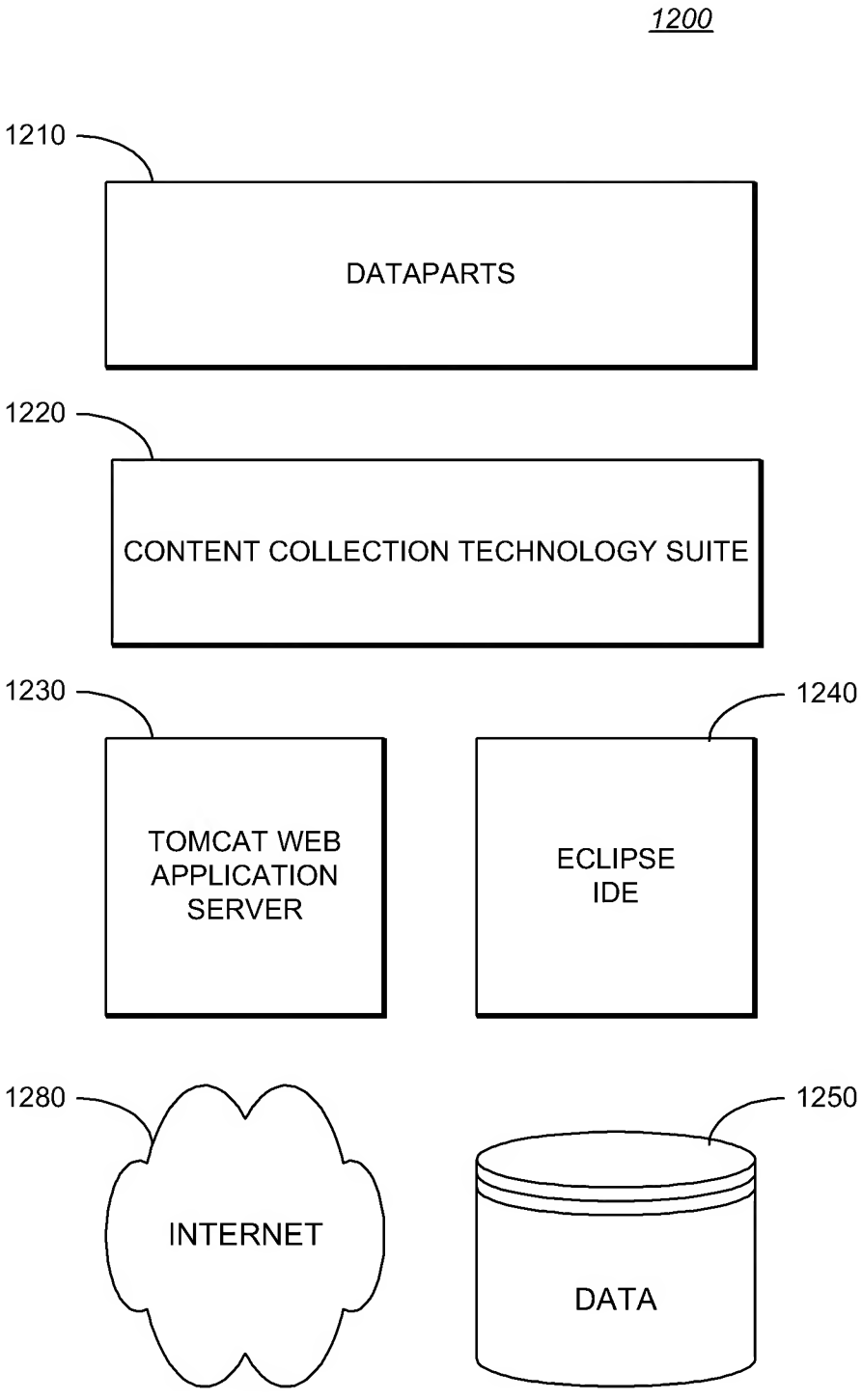
1200
1210
DATAPARTS
1220
CONTENT COLLECTION TECHNOLOGY SUITE
1230
TOMCAT WEB APPLICATION SERVER
1240
ECLIPSE IDE
1280
INTERNET
1250
DATA

FIG. 12

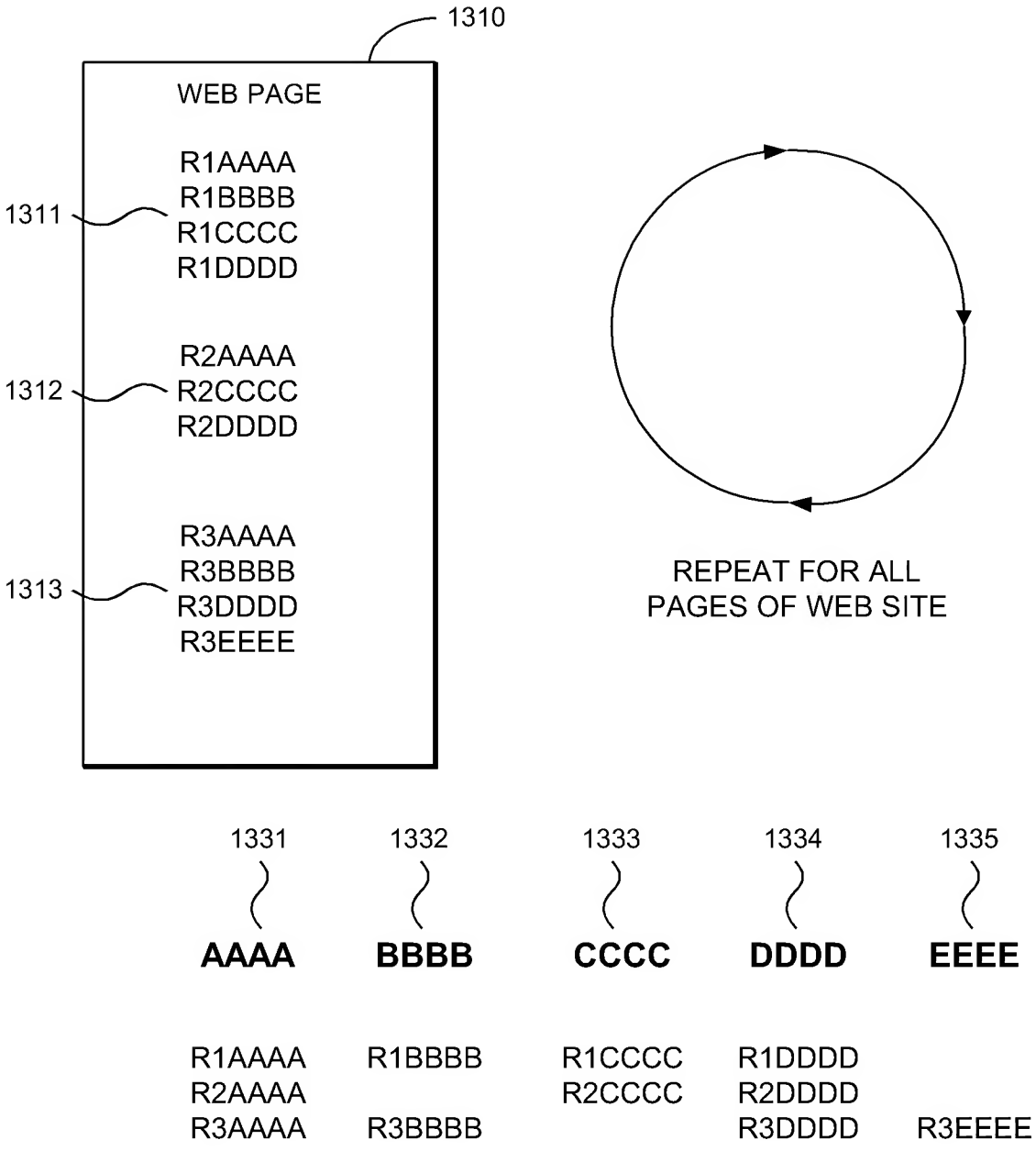
1310
WEB PAGE
R1AAAA
R1BBBB
1311
R1CCCC
R1DDDD
R2AAAA
1312
R2CCCC
R2DDDD
R3AAAA
R3BBBB
1313
R3DDDD
R3EEEE
REPEAT FOR ALL
PAGES OF WEB SITE
1331
1332
1333
1334
1335
AAAA
BBBB
CCCC
DDDD
EEEE
R1AAAA
R2AAAA
R3AAAA
R1BBBB
R3BBBB
R1CCCC
R2CCCC
R1DDDD
R2DDDD
R3DDDD
R3EEEE

FIG. 13

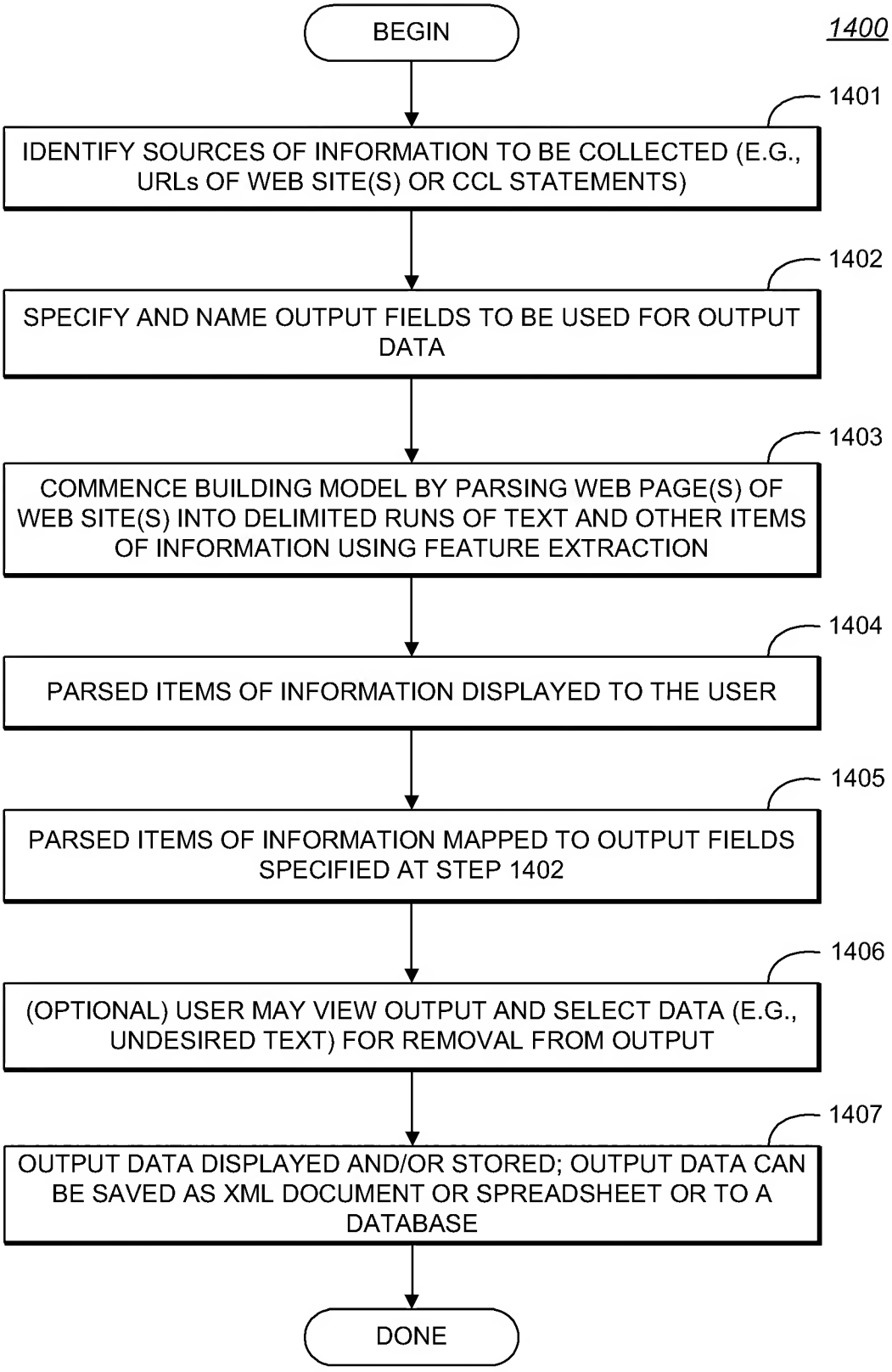
BEGIN
1400
1401
IDENTIFY SOURCES OF INFORMATION TO BE COLLECTED (E.G., URLs OF WEB SITE(S) OR CCL STATEMENTS)
1402
SPECIFY AND NAME OUTPUT FIELDS TO BE USED FOR OUTPUT DATA
1403
COMMENCE BUILDING MODEL BY PARSING WEB PAGE(S) OF WEB SITE(S) INTO DELIMITED RUNS OF TEXT AND OTHER ITEMS OF INFORMATION USING FEATURE EXTRACTION
1404
PARSED ITEMS OF INFORMATION DISPLAYED TO THE USER
1405
PARSED ITEMS OF INFORMATION MAPPED TO OUTPUT FIELDS SPECIFIED AT STEP 1402
1406
(OPTIONAL) USER MAY VIEW OUTPUT AND SELECT DATA (E.G., UNDESIRED TEXT) FOR REMOVAL FROM OUTPUT
1407
OUTPUT DATA DISPLAYED AND/OR STORED; OUTPUT DATA CAN BE SAVED AS XML DOCUMENT OR SPREADSHEET OR TO A DATABASE
DONE

FIG. 14